

# Survey data contamination through Large Language Models: Predicting LLM-generated answers to open narrative questions

Höhne<sup>1</sup>, Claassen<sup>1</sup>, Bach<sup>2</sup>, & Haensch<sup>3</sup>

<sup>1</sup> DZHW, Leibniz University Hannover

<sup>2</sup> University of Mannheim

<sup>3</sup> LMU Munich

**Current Innovations in Probability-based Household Internet Panel Research (CIPHER)**

Washington, DC (USA) – February 25 – 27, 2026

This research is funded by the  
German Society for Online Research



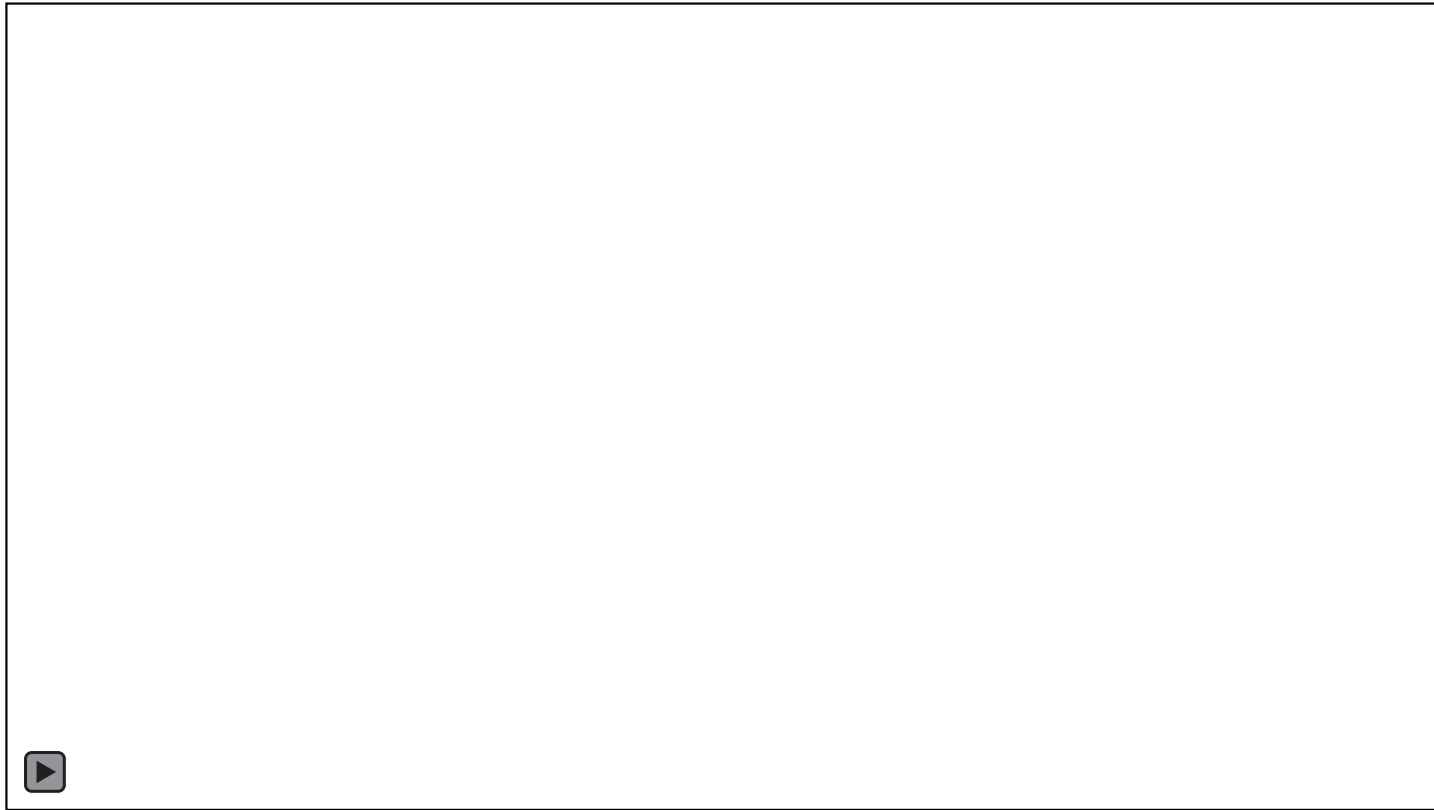
# Introduction I

- Growing demand for high-quality survey data (Knowledge Sourcing Intelligence 2023)
- Cost-efficient and streamlined web surveys replace other survey modes, especially in-person interviews (Schober 2018)
- Web surveys may not be suitable for primary survey mode
  - *Depressed response rates* (Daikeler et al. 2020)
  - *Frequently struggle with achieving high data quality* (Callegaro et al. 2015)
- No interviewers for assistance and to create trust, motivation, and engagement
  - *Respondents are on their own without monitoring* (Höhne et al. 2020)
  - *Web offers numerous opportunities to cut corners: so-called “cheating”* (Scott & Jerrit 2016)
  - *The advent of Large Language Models (LLMs) has fueled the problem further* (Rilla et al. 2025)

# Introduction II

- There is rumor about respondents prompting LLMs to answer open narrative questions
  - *Reducing response effort: formulating and entering answers is burdensome*
  - *Potential threat to the quality and integrity of survey outcomes*
  - *The extent of LLM-contaminated answers and how to detect them is unclear*
- For example, the Psychological Science journal just released that it ...
  - *... “demands an explicit statement on measures taken to reduce the risk of AI-generated responses for all online studies.”*
- In this study, we therefore address the following two research questions (RQs):
  - *What are the attributes of open narrative answers generated through LLMs? (RQ1)*
  - *Can we detect open narrative answers in web surveys generated through LLMs? (RQ2)*

# Showcase: Contamination through LLMs



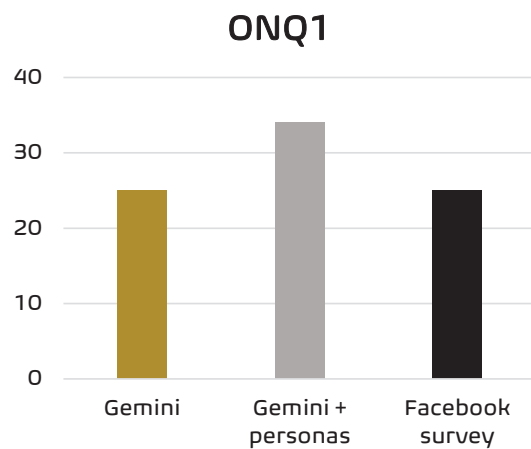
# Method: Data and Analyses

- Web survey on same-gender partnerships programmed with Unipark
  - *Three open narrative questions: Child adoption, discrimination, and final comment*
  - *For each question, we prompted Gemini 1.5 Pro (Google 2024) 800 times in February 2025*
  - *Gemini adopted personas – age, gender, education, and party preference – in 50% of the cases*
  - *We also conducted a web survey through Facebook (N = 1,512) in February/March 2024*
- RQ1: Text-as-data methods in the form of answer length and word choice
- RQ2: Predicting robotic language
  - *Fine-tuning BERT for each ONQ: LLM-generated text = “yes” or LLM-generated text = “unclear”*
  - *Performance evaluation: Precision, recall, and F1 score*

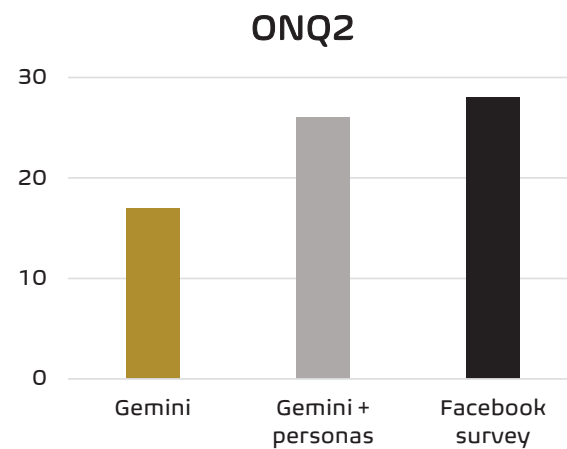
# Results: Exemplary Answers

Gemini	Gemini + personas	Facebook survey
<p>Jeder sollte die gleichen Chancen haben, eine Familie zu gründen. Liebe ist Liebe.</p> <p><b>Translation:</b> <i>Everyone should have the same opportunities to start a family. Love is love.</i></p>	<p>Ein Kind braucht 'ne Mutter und 'nen Vater. So is das nun mal vorgesehen.</p> <p><b>Translation:</b> <i>A child needs a mother and a father. That's how it's meant to be.</i></p>	<p>Hauptsache es wird sich gut um das Kind gekümmert.</p> <p><b>Translation:</b> <i>The most important thing is that the child is well taken care of.</i></p>

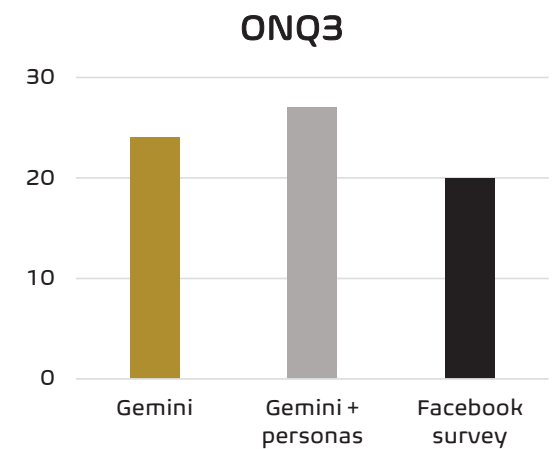
# Results: Answer Length (RQ1)



Note. Average number of words.  
One-way ANOVA:  $p < 0.001$ .



Note. Average number of words.  
One-way ANOVA:  $p < 0.001$ .



Note. Average number of words.  
One-way ANOVA:  $p < 0.001$ .



# Results: LLM-generated Text (RQ2)

Table 1. Prediction performance

	ONQ1	ONQ2	ONQ3
Training set size (60%)	960	960	758
Validation set size (20%)	320	320	253
Test set size (20%)	320	320	253
Precision	0.98	0.97	0.99
Recall	0.99	1.0	0.97
F1 score	0.98	0.99	0.98

Note. We used the “bert-base-german-cased” model via the “Simple Transformers” library in Python. For ONQ1 and ONQ2, we used all 800 Gemini answers as well as 800 randomly selected Facebook survey answers, respectively, to create a balanced sample. For ONQ3, in contrast, we used all 632 Facebook survey answers as well as 632 randomly selected Gemini answers.

# Discussion and Conclusion

- There are similarities between LLM-generated answers and those from the Facebook survey
  - *LLMs provide meaningful open narrative answers*
  - *No systematic differences regarding answer length*
  - *Word choice may offer clues when it comes to detecting LLM-generated answers*
- BERT reliably predicts LLM-generated answers
  - *Between 97 and 100 percent of the answers are correctly detected*
  - *Applies to answers from both Gemini and Gemini + personas*
- We currently explore further research possibilities
  - *Using BERT to predict prevalence of LLM-generated answers in web survey data*
  - *Making predictions based on closed questions*
  - *Examining other LLMs, such as GPT-4 and Llama 3.3*

Research Note

**Bots in web survey interviews: A showcase**

Jan Kareem Hühne<sup>1</sup> and Joshua Claassen<sup>2</sup>  
<sup>1</sup> Leibniz University Hannover, Germany  
<sup>2</sup> German Centre for Higher Education Research and Science Studies (DZHW), Germany

Sajjal Shaharia<sup>3</sup>  
<sup>3</sup> Otto von Guericke University Magdeburg, Germany

David Broneske<sup>4</sup>  
<sup>4</sup> German Centre for Higher Education Research and Science Studies (DZHW), Germany

**Abstract**  
 Cost- and time-efficient web surveys have progressively replaced other survey modes. These efficiencies can potentially cover the increasing demand for survey data. However, since web surveys suffer from low response rates, researchers and practitioners start considering social media platforms as new sources for respondent recruitment. Although these platforms provide advertisement and targeting systems, the data quality and integrity of web surveys recruited through social media might be threatened by bots. Bots have the potential to shift survey outcomes and thus political and social decisions. This is alarming since there is ample literature on bots and how they infiltrate social media platforms, distribute fake news, and possibly slow public opinion. In this study, we therefore investigate bot behavior in web surveys to provide new evidence on common wisdom about the capabilities of bots. We programmed four bots – two rule-based and two AI-based bots – and ran each bot  $N = 100$  times through a web survey on equal gender partnerships. We tested several bot prevention and detection measures, such as CAPTCHAs, invisible honey pot questions, and completion times. The results indicate that both rule- and AI-based bots come with impressive completion rates (up to 100%). In addition, we can prove conventional wisdom about bots in web surveys among CAPTCHAs and honey pot questions pose no challenges. However, there are clear differences between rule- and AI-based bots when it comes to web survey completion.

**Keywords**  
 rule-based bots, AI-based bots, web surveys, completion behavior, data integrity, data quality

**Corresponding author:**  
 Jan Kareem Hühne, Leibniz University Hannover, German Centre for Higher Education Research and Science Studies (DZHW), Research Infrastructure and Methods Division, Lange Laue 12, Hannover 30105, Germany  
 Email: [huehne@dzhw.de](mailto:huehne@dzhw.de)

International Journal of Market Research  
 Published online first  
 ISSN: 1473-2475  
 Sage



INTERNATIONAL JOURNAL OF SOCIAL RESEARCH METHODOLOGY  
 10.1080/15332673.2024.2316698

SHORT ARTICLE

**LLM-driven bot infiltration: protecting web surveys through prompt injections**

Jan Kareem Hühne<sup>1</sup>, Joshua Claassen<sup>2</sup> and Ben Lasse Wolf<sup>3</sup>  
<sup>1</sup> German Centre for Higher Education Research and Science Studies (DZHW), Leibniz University Hannover, Hannover, Germany

**Abstract**  
 Cost- and time-efficient web surveys potentially help covering the increasing survey data demand. However, since web surveys face low response rates, researchers consider social media platforms for recruitment. Although these platforms provide targeting tools, data quality and integrity might be threatened by bots. Established bot-detection are not reliable when it comes to LLM-driven bots linked to Large Language Models (LLMs). We therefore investigate whether and to what extent prompt injections help detecting LLM-driven bots in web surveys. We activated two LLM-driven bots with comparable abilities (LLM and LLaMA) to respond to an open-ended question. This question included no injection, a jailbreaking injection, or a prompt linking injection. Our results indicate that both bots react differently to prompt injections. While the less sophisticated LLM bot fails for the jailbreaking injection, the more sophisticated LLM bot fails for the prompt linking injection. This indicates that prompt injections must be tailored to bot sophistication.

**Keywords**  
 Data quality and integrity, jailbreaking injection, Large Language Models (LLMs), open-ended questions, prompt linking injection

**Article history**  
 Received 24 February 2023  
 Accepted 27 November 2023

**Introduction**  
 Web surveys have successively taken the place of other survey data collection methods, such as face-to-face interviews. Prominent social surveys, including the European Social Survey, have adopted web survey data collection. Due to their cost- and time-efficiency, web surveys are seen as a strong contender meeting the high survey data demand (Knowledge Sourcing Intelligence, 2023). Nonetheless, they may not be prepared to replace other data collection methods, as they result in low response rates (Dauker et al., 2020).

Researchers explore alternative ways of recruiting, including social media platforms, such as Facebook, which provide targeting tools (Zinski, 2023). While social media recruitment offers access to a vast respondent pool, data quality and integrity of such surveys face risks from bots. Bots are automated programs designed to interact with web-based systems (Griffin et al., 2022; Hühne et al., 2023; Storozak

**Contact** Jan Kareem Hühne [huehne@dzhw.de](mailto:huehne@dzhw.de), German Centre for Higher Education Research and Science Studies (DZHW), Leibniz University Hannover, Lange Laue 12, Hannover 30105, Germany  
 Email: [huehne@dzhw.de](mailto:huehne@dzhw.de)

**Supplemental data** for this article can be accessed online at <https://doi.org/10.1080/15332673.2024.2316698>.

**©** 2024 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.  
 This is an open access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The terms on which this article has been published allow for the copying of the full text for personal use, provided the article is properly cited.



Report

**Identifying Bots Through LLM-Generated Text in Open Narrative Responses: A Proof-of-Concept Study**

Joshua Claassen<sup>1</sup>, Jan Kareem Hühne<sup>1</sup>, Ruben Bach<sup>2</sup>, and Anna-Carolina Haensch<sup>3</sup>

**Abstract**  
 Online survey participants are frequently recruited through social media platforms, open online access panels, and river sampling approaches. Such online surveys are threatened by bots that limit survey outcomes and exploit incentives. In this proof-of-concept study, we advance the identification of bots driven by Large Language Models (LLMs) through the prediction of LLM-generated text in open narrative responses. We conducted an online survey on same-gender partnerships, including three open narrative questions, and recruited 1512 participants through Facebook. In addition, we visited two LLM-driven bots, each of which responded to the open narrative questions 400 times. Open narrative responses synthesized by our bots were labeled as containing LLM-generated text ("yes"). Facebook responses were assigned a proxy label ("unclear") as they may contain bots themselves. Using this binary label as ground truth, we fine-tuned predictor models relying on the Bidirectional Encoder Representations from Transformers (BERT) model, resulting in an impressive prediction performance. The models accurately identified between 97% and 100% of bot responses. However, prediction performance decreased if the models make predictions about questions they were not fine-tuned with. Our study contributes to the ongoing discussion on bots and extends the methodological toolkit for protecting the quality and integrity of online survey data.

**Keywords**  
 LLM-driven bots, Data quality and integrity, Large Language Models (LLMs), Machine learning, Response behavior, Web surveys, Explainable AI

**Research Infrastructure and Methods Division, German Centre for Higher Education Research and Science Studies (DZHW), Leibniz University Hannover, Hannover, Germany**  
<sup>2</sup>German Centre for European Social Research (GZES), University of Mannheim, Mannheim, Germany  
<sup>3</sup>Open Data Science and AI Lab, Ludwig Maximilians University (LMU), Munich, Germany

**Corresponding Author:**  
 Joshua Claassen, Research Infrastructure and Methods Division, German Centre for Higher Education Research and Science Studies (DZHW), Leibniz University Hannover, Lange Laue 12, Hannover 30105, Germany  
 Email: [claassen@dzhw.de](mailto:claassen@dzhw.de)



Many thanks for your attention!

[hoehne@dzhw.eu](mailto:hoehne@dzhw.eu)

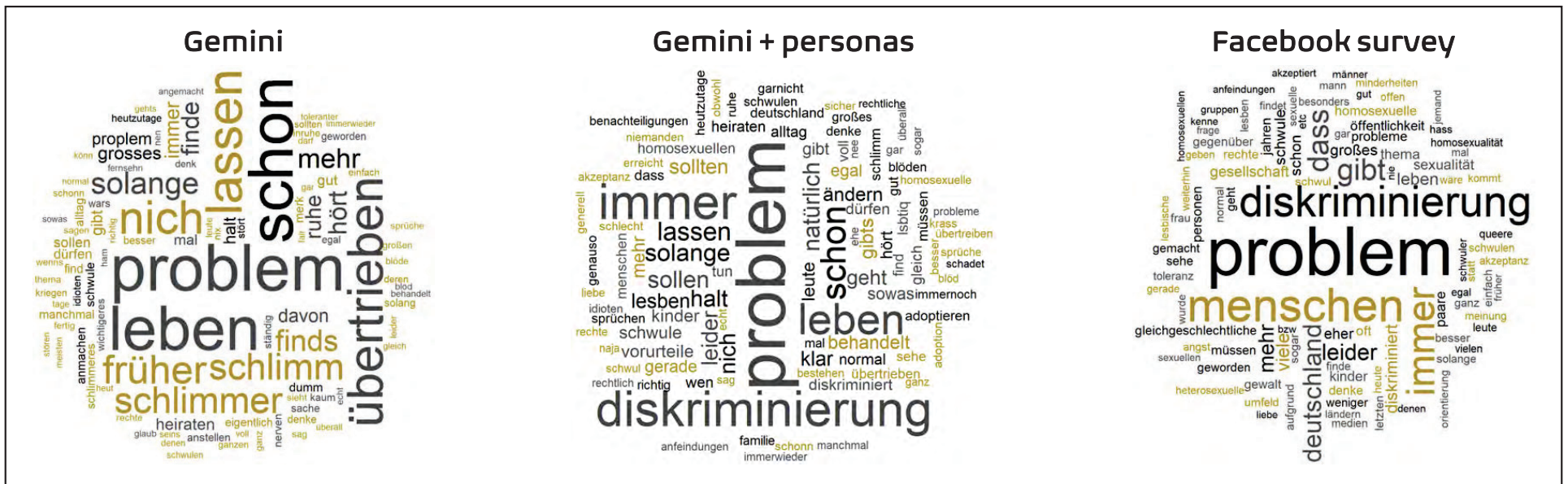
# Literature

- Callegaro, M., Lozar Manfreda, K., & Vehovar, V. (2015). Web survey methodology. Sage. <https://doi.org/10.4135/9781529799651>
- Clifford, S., & Jerit, J. (2016). Cheating on political knowledge questions in online surveys: An assessment of the problem and solutions. *Public Opinion Quarterly*, 80, 858–887. <https://doi.org/10.1093/poq/nfw030>
- Daikeler, J., Bosnjak, M., & Lozar Manfreda, K. (2020). Web versus other survey modes: An updated and extended meta-analysis comparing response rates. *Journal of Survey Statistics and Methodology*, 8, 513–539. <https://doi.org/10.1093/jssam/smz008>
- Google. (2024). Gemini: A family of highly capable multimodal models. arXiv. <https://doi.org/10.48550/arXiv.2312.11805>
- Höhne, J.K., Cornesse, C., Schlosser, S., Couper, M.P., & Blom, A. (2020). Looking up answers to political knowledge questions in web surveys. *Public Opinion Quarterly*, 84, 986–999. <https://doi.org/10.1093/poq/nfaa049>
- Knowledge Sourcing Intelligence (2023). Global online survey software market size, share, opportunities, COVID 19 impact, and trends by application, by product, and by geography – forecasts from 2023 to 2028. <https://www.knowledge-sourcing.com/report/global-online-survey-software-market>
- Rilla, R., Werner, T., Yakura, H., Rahwan, I., & Nussberger, A.-M. (2025). Recognising, anticipating, and mitigating LLM pollution of online behavioural research. arXiv. <https://doi.org/10.48550/arXiv.2508.01390>
- Schober, M. F. (2018). The future of face-to-face interviewing. *Quality Assurance in Education*, 26, 290–302. <https://doi.org/10.1108/QAE-06-2017-0033>

# Appendix A: Open Narrative Questions

- **ONQ1:** In the last question, you indicated to find it (very good | rather good | rather not good | not good at all) that married same-sex partners in Germany can adopt children. Please explain to us in your own words why you chose this response.
- **ONQ2:** In your opinion, to what extent is discrimination against gay, lesbian and bisexual people a problem or no problem in Germany?
- **ONQ3:** Finally, we would like to give you the opportunity to say something about our survey. Do you have any comments or suggestions on the survey as a whole or on individual questions?

# Appendix B: Word Choice (ONQ2)



Note. Each word cloud contains the 100 most frequently mentioned words (ONQ1) among Gemini, Gemini + personas, and the Facebook survey, respectively. The size of a word is proportional to its frequency.



# Appendix D: Prompt (Baseline Design)

## *LLM version*

gemini-1.5-pro-002.

## *Open narrative questions – Prompt design (Gemini)*

Verhalte dich wie eine Person, die an einer Umfrage teilnimmt, und schreibe eine Antwort auf Deutsch basierend auf deren Denkweise/Eigenschaften für die folgende Frage: {question}  
Gib eine kurze und prägnante Antwort.

## *Open narrative questions – Prompt design (Gemini + personas)*

Verhalte dich wie eine {age} Jahre alte deutschsprachige {gender} Person mit {education} und {party preference} nahestehend, die an einer Umfrage teilnimmt, und schreibe eine Antwort auf Deutsch basierend auf deren Denkweise/Eigenschaften für die folgende Frage: {question}  
Gib eine kurze und prägnante Antwort.  
Berücksichtige dabei deine bisherigen Antworten: {history}"

## *Personas (Gemini + personas)*

Age: 18 to 89 years

Gender: female or male

Education: low education, medium education, or high education

Party preference: SPD, CDU/CSU, Greens, FDP, AfD, or Left

## *Gemini parameters*

generation\_config = {"temperature": 1.0, "max\_output\_tokens": 2048}.

# Appendix E: Token Analysis

LLM-generated text = "yes"				LLM-generated text = "unclear"		
	Token	Attribution score	Frequency	Token	Attribution score	Frequency
ONQ1	(1) Fin	0.78	126	(1) auch	0.25	30
	(2) ##d	0.52	111	(2) Kinder	0.20	71
	(3) is	0.20	38	(3) Eltern	0.19	38
	(4) Ein	0.19	28	(4) und	0.17	92
	(5) ich	0.16	140	(5) zu	0.17	37
ONQ2	(1) schon	0.59	71	(1) Problem	0.31	96
	(2) Is	0.49	35	(2) nicht	0.23	73
	(3) doch	0.42	43	(3) oder	0.22	31
	(4) is	0.39	27	(4) wird	0.21	40
	(5) Also	0.39	43	(5) werden	0.20	36
ONQ3	(1) Also	0.47	46	(1) der	0.20	48
	(2) verständlich	0.43	30	(2) es	0.16	34
	(3) waren	0.27	44	(3) ##en	0.16	31
	(4) Fragen	0.25	72	(4) nicht	0.15	47
	(5) Die	0.24	39	(5) den	0.15	26